University of Thessaly
School of Medicine
Research Methodology in Biomedicine, Biostatistics and Clinical Bioinformatics

Multicollinearity: diagnostics and PCA as a method of handling

Πολυσυγγραμικότητα: διαγνωστικές μέθοδοι και Ανάλυση Κύριων Συνιστωσών ως μέθοδος χειρισμού

Scientific Committee:
Apostolos Batsidis, *MSc, PhD, Assistant Professor, Probability, Statistics and Operations Research Unit, Department of Mathematics, University of Ioannina* (Supervisor)
Ioannis Stefanidis, *MD, PhD, Professor of Internal Medicine/Nephrology, Faculty of Medicine, University of Thessaly*
Chrysoula Doxani, *MSc, MD, PhD, Research Fellow in Genetic Pharmacoepidemiology, University of Thessaly*

Paraskevi Botsi
Email: pbotsi@med.uth.gr

September 2017

Abstract

The present Master's thesis seeks to develop a better understanding of multicollinearity, which is present when there is a strong correlation between independent variables in multiple linear regression analysis. In this frame, we will first describe methods to diagnose the problem of multicollinearity and then we will describe Principal Component Analysis[1] as a method of handling of this adverse situation. Finally, the method is performed in a data set related with cystic fibrosis.

In this context the structure of the dissertation is as follows. Chapter 1 'Introduction' analyses the concept of multicollinearity, indicates the collinearity diagnostic indexes, as well the problems that creates in multiple regression analysis. Afterwards, the PCA which is a method of handling multicollinearity is introduced.

Chapter 2 'Methods & Results', PCA method is implemented in a data set, the collinearity indications are detected and the results from the correction procedure by applying PCA are presented.

Finally, in Chapter 3 'Conclusions' findings are summarized and the importance of detecting and dealing with multicollinearity in multiple regression analysis is pointed out.

Key words
Multicollinearity, Regression Analysis, Principal Component Analysis

[1] Principal Component Analysis, PCA

Περίληψη

Στην εργασία αυτή, αναλύεται η πολυσυγγραμικότητα, η οποία δημιουργείται στην πολλαπλή γραμμική παλινδρόμηση όταν υπάρχει ισχυρή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Αρχικά γίνεται περιγραφή των μεθόδων εντοπισμού της πολυσυγγραμικότητας και στη συνέχεια αναλύεται ως μέθοδος χειρισμού και αντιμετώπισης αυτής της ανεπιθύμητης κατάστασης, η Ανάλυση Κύριων Συνιστωσών[2]. Τέλος εφαρμόζεται σε ένα σύνολο δεδομένων που σχετίζονται με την κυστική ίνωση.

Στο πλαίσιο αυτό, η διάρθρωση της διπλωματικής διατριβής είναι η ακόλουθη. Στο Κεφάλαιο 1 'Εισαγωγή' (Introduction) αναλύεται η έννοια της συγγραμικότητας, επισημαίνονται οι δείκτες εντοπισμού της πολυσυγγραμικότητας καθώς και τα προβλήματα που δημιουργεί στην πολυμεταβλητή ανάλυση παλινδρόμησης. Έπειτα, γίνεται μια σύντομη εισαγωγή στη μέθοδο της Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis), η οποία χρησιμοποιείται μεταξύ άλλων ως μέθοδος χειρισμού του προβλήματος της πολυσυγγραμμικότητας.

Στο Κεφάλαιο 2 'Μέθοδοι – Αποτελέσματα' (Methods – Results) γίνεται εφαρμογή της μεθόδου ΑΚΣ σε ένα σύνολο δεδομένων, εντοπίζονται οι ενδείξεις της πολυσυγγραμικότητας και παρουσιάζονται τα αποτελέσματα διόρθωσης αυτής μετά την εφαρμογή της ΑΚΣ.

Τέλος στο Κεφάλαιο 3 'Συμπεράσματα', συνοψίζονται τα αποτελέσματα και επισημαίνεται η σπουδαιότητα ανίχνευσης και αντιμετώπισης της πολυσυγγραμικότητας στην ανάλυση πολλαπλής παλιδρόμησης.

Λέξεις κλειδιά
Πολυσυγγραμικότητα, Ανάλυση Παλινδρόμησης, Ανάλυση Κύριων Συνιστωσών

[2] Ανάλυση Κύριων Συνιστωσών, ΑΚΣ

Chapter 1: Introduction

One of the important assumptions in multiple regression analysis is that the independent variables are not strongly interrelated because otherwise the interpretation of a model may not be valid when linear relationship exists among predictor variables. Collinearity is an adverse situation when this assumption is not met i.e. there are two covariates which are moderately or highly correlated and when there are more than two, this is multicollinearity. False-positive results (Type I error) and false-negative results (Type II error) may be present when regression coefficients are biased by collinearity.[1][2] There are two types of multicollinearity: Structural multicollinearity which caused by creating a new independent variable from other predictors and data-based multicollinearity, caused mainly by observational experiments or by the inability to interfere in the collection of data. The multicollinearity can be caused by insufficient data collection, dummy variables, or simply by including in regression analysis an independent variable which is actually a linear combination of other variables.[3] There are numerous indicators to suggest that collinearity exists, such as:

- It is possible and due to multicollinearity, none of the independent variables are not statistically significant, *P-value > 0.05*, while the F test is highly statistically significant.
- Variance Inflation Factor: Variables with large VIF, $max\text{VIF}_j$ >10

  $\text{VIF}_j = \frac{1}{1-R^2_j}$

  R^2_j: is the coefficient of determination of a regression of explanator j on all the other explanators.
- Tolerance: When its value is less than 0.1

  Tolerance = $1 - R^2_j$
- Condition number: A condition number greater than 1000

  Condition number is the ratio of the largest Eigenvalue to smallest Eigenvalue of matrix X.
- Condition index: A condition index greater than 15 indicates a possible problem and a condition index greater than 30 suggests a serious problem with collinearity.

  Condition index is the square root of ratio of the largest Eigenvalue to each *j*th Eigenvalue of matrix X.
- Variance-decomposition proportion: At least two regression coefficients with variance-decomposition proportion greater than 0.5. The variance-decomposition proportion for the *j*th regression coefficient associated with the *i*th component is defined as $\pi ij = \Phi ij / \Phi j$, where $\Phi ij = v^2 ij / \lambda i$, $\Phi j = \sum_{i=1}^{p} \Phi ij$.

- The absolute value of the correlation coefficient between two independent variables in the correlation matrix. The Pearson correlation coefficient indicates positive linear correlation when r = +1, no linear correlation when r = 0, and negative linear correlation when r = -1.[4]

After identifying the presence of collinearity in a data set, there are several methods for dealing with multicollinearity, such as Ridge Regression, Partial Least Squares Regression and Principal Component Analysis. In this thesis we are going to analyze further the last method.

'Principal component analysis is a widely used multivariate statistical method, which can transform the original variables into a set of new orthogonal variables, so that most information is contained in the first few components with the largest variance'.[5] The number of principal components may be less or equal to the number of the original variables. Karl Pearson invented PCA method in 1901 and it was later independently developed and named by Harold Hotelling in the 1930s.[6] The main goal of PCA was defined by Karl Pearson in his paper: 'In many physical, statistical and biological investigations it is desirable to represent a system of points in plane, three or higher demensioned space by the 'best fitting' straight line or plane.'[7]

Besides dealing with multicollinearity, Principal Components Analysis, as an exploratory multivariate statistical technique, capable for dimension reduction, is widely used in image compression, neuroscience, gene expression and other applications in many multidisciplinary fields.

The main aim of this thesis is to detect multicollinearity between highly correlated independent variables, when performing multivariate analysis in a data set and apply PCA as a method of solving multicollinearity's problem by using the statistical software SPSS 23.

Chapter 2: Methods - Results

The set of data was obtained from the book 'Biostatistics for Medical and Biomedical Practitioners' by Julien Hoffman.[8] The data set includes a number of variables from patients with cystic fibrosis. These variables were considered most likely to predict maximal static expiratory pressure PEmax, a measure of malnutrition and can be tested to determine the most useful explanatory or predictive model. PEmax *(Y)* is the dependent variable and the independent variables are Age *(X_1)*, Sex, Height *(X_2)*, Weight *(X_3)*, Body Mass *(X_4)*, Forced Expiratory Volume *(X_5)*, Residual Volume *(X_6)*, Functional Residual Capacity *(X_7)* and Total Lung Capacity *(X_8)*. We exclude Sex, which is a categorical variable from the independent variables.

We are going to formulate a predictive model for PEmax from the set of the independent variables. We proceed with regression analysis with the dependent variable Y and all independent variables X. It is possible and due to multicollinearity, none of the independent variables are not statistically significant, *P-value > 0.05*. Table 1 provides information about the multiple regression model. The Adjusted R^2 of our model is 0.509 with $R^2 = 0.673$. This means that the linear regression explains 67.3% of the variance in the data. R^2 could be affected by the number of independent variables and sample size, therefore the Adjusted R^2 is a better index for comparing the goodness of fit between linear models since it is designed to compensate for the optimistic bias of R^2.[9]

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,820[a]	**,673**	**,509**	23,43149	,673	4,109	8	16	**,008**	2,378

a. Predictors: (Constant), total lung capacity, body mass (% of normal), age in years, forced expiratory volume, residual volume, height (cm), functional residual capacity, weight (kg)

b. Dependent Variable: maximum expiratory pressure

Table 1: Multiple Regression Model Summary

The linear regression's F-test has the null hypothesis that the model explains zero variance in the dependent variable. The F-test is highly statistically significant with *P-value = 0.008 < 0.05*, thus we can assume that the model explains a significant amount of the variance in maximum expiratory pressure.

In Table 2, we detect that none of the independent variables seems to be statistical significant *P-value > 0.05*, nevertheless the regression model is statistical significant. This is an indication that collinearity exists among the predictor variables. The indexes Tolerance and VIF

show us the magnititude of multicollinearity, since 3 of 9 independent variables are strongly correlated.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	151,183	187,490		,806	**,432**	-246,278	548,644					
	age in years	-3,339	4,202	-,505	-,795	**,438**	-12,246	5,568	,613	-,195	-,114	**,051**	**19,751**
	height (cm)	-,350	,808	-,225	-,433	**,671**	-2,064	1,363	,599	-,108	-,062	**,076**	**13,203**
	weight (kg)	3,179	1,759	1,702	1,807	**,090**	-,551	6,909	,635	,412	,258	**,023**	**43,345**
	body mass (% of normal)	-1,768	1,034	-,635	-1,710	**,107**	-3,960	,424	,230	-,393	-,245	,148	6,737
	forced expiratory volume	1,352	,702	,453	1,925	**,072**	-,137	2,841	,453	,434	,275	,370	2,703
	residual volume	,248	,145	,637	1,718	**,105**	-,058	,555	-,284	,395	,246	,149	6,710
	functional residual capacity	-,309	,326	-,404	-,948	**,357**	-,999	,382	-,417	-,231	-,136	,113	8,866
	total lung capacity	,113	,440	,057	,256	**,801**	-,821	1,046	-,182	,064	,037	,410	2,440

a. Dependent Variable: maximum expiratory pressure

Table 2: Collinearity Statistics and significant values of Independent Variables coefficients

In Table 3, the Eigenvalues and Condition Indexes are presented. Eigenvalues indicate how many distinct dimensions there are among independent variables and the variables are highly intercorrelated when several Eigenvalues are close to 0, then the matrix is said to be ill-conditioned.[8] Condition Index values greater than 30, also reveals a severe problem of multicollinearity.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions								
				(Constant)	age in years	height (cm)	weight (kg)	body mass (% of normal)	forced expiratory volume	residual volume	functional residual capacity	total lung capacity
1	1	8,464	1,000	,00	,00	,00	,00	,00	,00	,00	,00	,00
	2	,391	4,652	,00	,00	,00	,00	,00	,01	,01	,00	,00
	3	,091	9,660	,00	,01	,00	,00	,00	,22	,01	,00	,00
	4	,023	19,235	,00	,05	,00	,03	,04	,12	,08	,04	,01
	5	,014	25,019	,01	,01	,01	,05	,00	,26	,24	,02	,13
	6	,010	29,015	,01	,00	,00	,02	,09	,10	,02	,00	,42
	7	**,005**	**42,575**	,00	,04	,01	,01	,08	,11	,60	,86	,17
	8	**,003**	**57,341**	,00	,61	,28	,08	,14	,01	,03	,04	,04
	9	**,000**	**147,280**	,99	,28	,70	,80	,65	,17	,00	,03	,23

a. Dependent Variable: maximum expiratory pressure

Table 3: Collinearity Diagnostics

The Scree plot diagram shows the Eigenvalues on y-axis and the number of factors-components on x-axis.

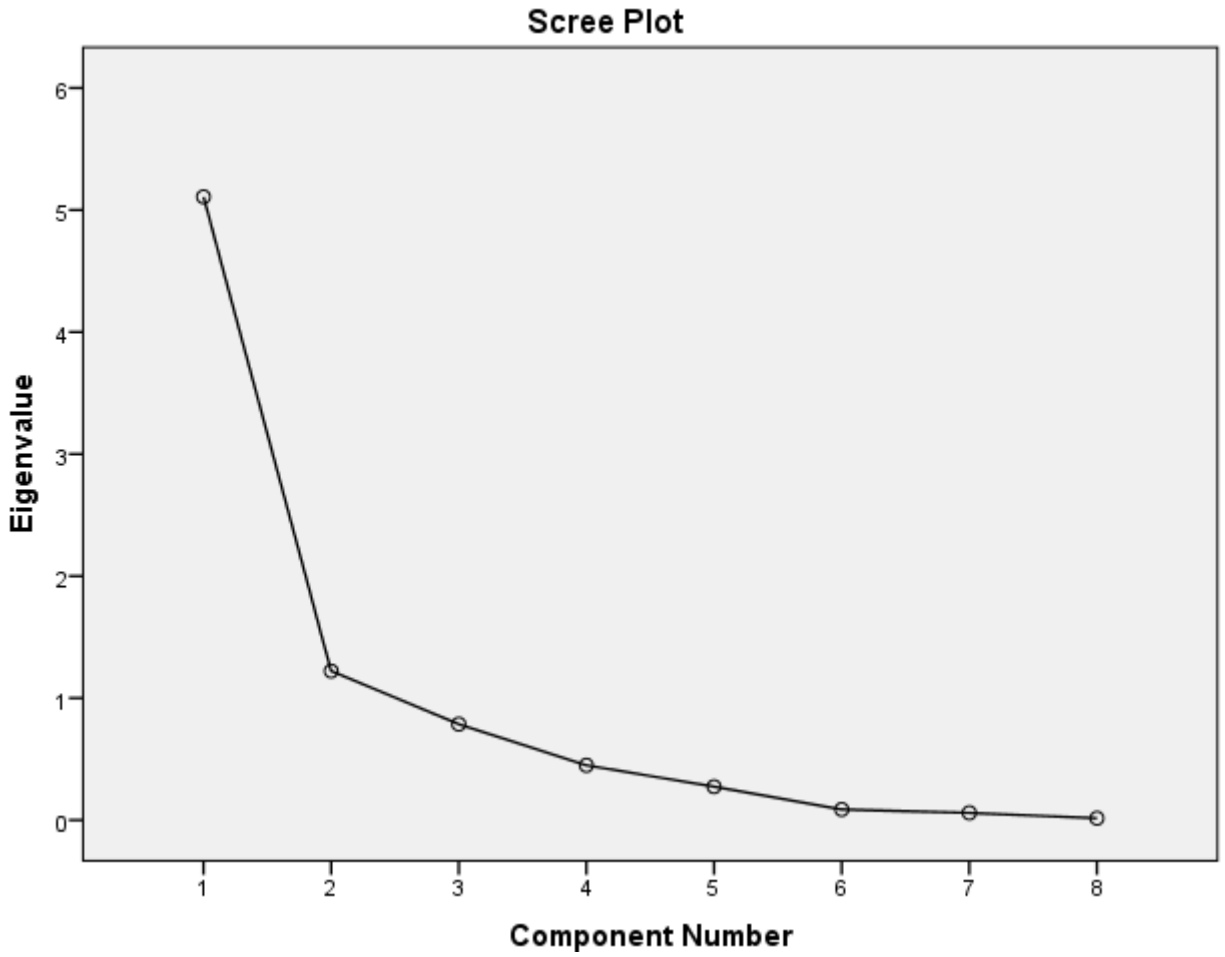

Diagram 1: Scree Plot

Finally, by extracting the Pearson's correlation coefficient r, for all the independent variables, we can observe in Table 4, that there are plenty of statistically significant correlated independent variables, *P-value < 0.05*.

Correlations

		age in years	height (cm)	weight (kg)	body mass (% of normal)	forced expiratory volume	residual volume	functional residual capacity	total lung capacity
age in years	Pearson Correlation	1	**,926****	**,906****	,378	,294	**-,532****	**-,639****	**-,469***
	Sig. (2-tailed)		**,000**	**,000**	,063	,153	**,006**	**,001**	**,018**
	N	25	25	25	25	25	25	25	25
height (cm)	Pearson Correlation	,926**	1	**,921****	**,441***	,317	**-,571****	**-,624****	**-,457***
	Sig. (2-tailed)	,000		**,000**	**,027**	,123	**,003**	**,001**	**,022**
	N	25	25	25	25	25	25	25	25
weight (kg)	Pearson Correlation	,906**	,921**	1	**,673****	**,449***	**-,633****	**-,617****	**-,418***
	Sig. (2-tailed)	,000	,000		**,000**	**,024**	**,001**	**,001**	**,037**
	N	25	25	25	25	25	25	25	25
body mass (% of normal)	Pearson Correlation	,378	,441*	,673**	1	**,546****	**-,615****	**-,434***	-,365
	Sig. (2-tailed)	,063	,027	,000		**,005**	**,001**	**,030**	,073
	N	25	25	25	25	25	25	25	25
forced expiratory volume	Pearson Correlation	,294	,317	,449*	,546**	1	**-,691****	**-,665****	**-,443***
	Sig. (2-tailed)	,153	,123	,024	,005		**,000**	**,000**	**,027**
	N	25	25	25	25	25	25	25	25
residual volume	Pearson Correlation	-,532**	-,571**	-,633**	-,615**	-,691**	1	**,877****	**,607****
	Sig. (2-tailed)	,006	,003	,001	,001	,000		**,000**	**,001**
	N	25	25	25	25	25	25	25	25
functional residual capacity	Pearson Correlation	-,639**	-,624**	-,617**	-,434*	-,665**	,877**	1	**,704****
	Sig. (2-tailed)	,001	,001	,001	,030	,000	,000		**,000**
	N	25	25	25	25	25	25	25	25
total lung capacity	Pearson Correlation	-,469*	-,457*	-,418*	-,365	-,443*	,607**	,704**	1
	Sig. (2-tailed)	,018	,022	,037	,073	,027	,001	,000	
	N	25	25	25	25	25	25	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4: Correlation Matrix between Independent Variables

From the previous steps, it is confirmed that multicollinearity is present and the next step is to perform Principal Component Analysis as a method of handling it. Initially we begin by saving the standardized values from all the variables and also the Mean value of each dependent and independent variable.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
age in years	25	7,00	23,00	**14,4800**	5,05899	25,593
height (cm)	25	109,00	180,00	**152,8000**	21,50000	462,250
weight (kg)	25	12,90	73,80	**38,4040**	17,89813	320,343
body mass (% of normal)	25	64,000	97,000	**78,28000**	12,005277	144,127
forced expiratory volume	25	18,00	57,00	**34,7200**	11,19717	125,377
residual volume	25	158,00	449,00	**258,8000**	85,66748	7338,917
functional residual capacity	25	104,00	268,00	**155,4000**	43,71880	1911,333
total lung capacity	25	81,00	147,00	**114,0000**	16,96811	287,917
maximum expiratory pressure	25	65,00	195,00	**109,1200**	33,43691	1118,027
Valid N (listwise)	25					

Table 5: Descriptive Statistics of Dependent and Independent Variables

Afterward, we select, Dimension Reduction → Factor, then enter the standardized independent variables and in Factor Analysis: Extraction dialog box click on Method → Principal Components and Factors to extract → 8, as the number of our independent variables. In Factor Analysis: Factor Scores dialog box select 'Save as variables', Method → Regression. In Factor

Analysis: Factor Descriptives dialog box select 'KMO and Bartlett's test of sphericity'. The SPSS generates the 8 Regression Factors.

From Table 6, it is confirmed that principal component analysis is a suitable method for handling multicollinearity since Kaiser Meyer Olkin (KMO) statistical tool, KMO = 0.709 > 0.6. The KMO index ranges from 0 to 1 and when KMO is greater than 0.6, the data is appropriate for factor analysis.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**,709**
Bartlett's Test of Sphericity	Approx. Chi-Square	204,623
	df	28
	Sig.	,000

Table 6: KMO and Bartlett's Test of Sphericity

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,108	63,850	63,850	5,108	63,850	63,850
2	1,222	15,269	79,119	1,222	15,269	79,119
3	,786	9,829	88,948	,786	9,829	88,948
4	,449	5,615	94,563	,449	5,615	94,563
5	,274	3,423	97,986	,274	3,423	97,986
6	,087	1,088	99,075	,087	1,088	99,075
7	,059	,739	99,814	,059	,739	99,814
8	,015	,186	100,000	,015	,186	100,000

Extraction Method: Principal Component Analysis.

Table 7: The Eigenvalue and Percentage of Variance for each Principal Component

Component Matrix[a]

	Component							
	1	2	3	4	5	6	7	8
Zscore: age in years	,826	,514	-,115	,074	,070	-,114	-,123	,055
Zscore: height (cm)	,842	,492	-,037	,053	,030	,155	,135	,032
Zscore: weight (kg)	,892	,374	,225	,003	,049	-,003	-,034	-,097
Zscore: body mass (% of normal)	,686	-,183	,608	-,343	-,055	-,065	,036	,032
Zscore: forced expiratory volume	,675	-,558	,176	,310	,325	,029	-,010	,009
Zscore: residual volume	-,867	,327	,022	-,094	,326	-,127	,100	-,006
Zscore: functional residual capacity	-,876	,238	,292	-,161	,156	,165	-,114	,014
Zscore: total lung capacity	-,691	,259	,485	,439	-,161	-,037	,020	,009

Extraction Method: Principal Component Analysis.

a. 8 components extracted

Table 8: Component Matrix of Standardized Independent Variables

From Table 8 the form of each principal component is given. To be more specific:

$$\boldsymbol{C_1 = 0.826*X'_1 + 0.842*X'_2 + 0.892*X'_3 + 0.686*X'_4 + 0.675*X'_5 - 0.867*X'_6 - 0.876*X'_7 - 0.691*X'_8}$$

$$C_2 = 0.514*X'_1 + 0.492*X'_2 + 0.374*X'_3 - 0.183*X'_4 + 0.558*X'_5 + 0.327*X'_6 + 0.238*X'_7 + 0.259*X'_8$$

$$C_3 = -0.115*X'_1 - 0.037*X'_2 + 0.225*X'_3 + 0.608*X'_4 + 0.176*X'_5 + 0.022*X'_6 + 0.292*X'_7 + 0.485*X'_8$$

$$\boldsymbol{C_4 = 0.074*X'_1 + 0.053*X'_2 + 0.003*X'_3 - 0.343*X'_4 + 0.310*X'_5 - 0.094*X'_6 - 0.161*X'_7 + 0.439*X'_8}$$

$$\boldsymbol{C_5 = 0.070*X'_1 + 0.030*X'_2 + 0.049*X'_3 - 0.055*X'_4 + 0.325*X'_5 + 0.326*X'_6 + 0.156*X'_7 - 0.161*X'_8}$$

$$C_6 = -0.114* X'_1 + 0.155* X'_2 - 0.003* X'_3 - 0.065* X'_4 + 0.029* X'_5 - 0.127* X'_6 + 0.165* X'_7 - 0.037* X'_8$$

$$C_7 = -0.123* X'_1 + 0.135* X'_2 - 0.034* X'_3 + 0.036* X'_4 - 0.010* X'_5 + 0.100* X'_6 - 0.114* X'_7 + 0.020* X'_8$$

$$C_8 = 0.055* X'_1 + 0.032* X'_2 - 0.097* X'_3 + 0.032* X'_4 + 0.009* X'_5 - 0.006* X'_6 + 0.014* X'_7 + 0.009* X'_8$$

where $X'_1, X'_2, X'_3, X'_4, X'_5, X'_6, X'_7, X'_8$ denotes the standardized values (Zscore) of the respective independent variable.

At the last step, we conduct regression analysis, excluding the constant in equations, with dependent variable the standardized Y value and as independent variables the regression factors that have been generated in the previous step.

Table 9 displays the Adjusted R^2 and the statistical significance of each model and by comparing the values of Adjusted R^2, we conclude that model 5 has the largest Adjusted R^2 value (0.516) and the smallest Standard Error of Estimate (0.68). Its F value is equal to 6.621 which is statistical significant with *P-value = 0.018 < 0.05*.

Model Summary[j,k]

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,544[a]	,296	,267	,83912279	,296	10,085	1	24	,004	
2	,601[c]	,362	,306	,81606880	,066	2,375	1	23	,137	
3	,605[d]	,366	,280	,83140792	,005	,159	1	22	,694	
4	,696[e]	,484	,386	,76773060	,118	4,801	1	21	,040	
5	**,783[f]**	**,613**	**,516**	**,68187356**	**,128**	**6,621**	**1**	**20**	**,018**	
6	,787[g]	,619	,498	,69401795	,006	,306	1	19	,586	
7	,789[h]	,622	,475	,71001423	,003	,154	1	18	,700	
8	,820[i]	,673	,519	,67984417	,051	2,633	1	17	,123	2,378

Table 9: Model Summary

From Table 10, we extract the Standardized Coefficients for model 5 and the Eq. (1) is

$$\widehat{y}'_i = \sum B'_i C_i \quad \textbf{(1)}$$

After taking into consideration the largest Adjusted R^2, the smallest Standard Error of Estimate and the significance level of each factor, we are prepared for the formation of our model. Thus, the equation Eq. (1) is the following:

$$\widehat{y}'_5 = 0.544*C_1 + 0.343*C_4 + 0.358*C_5 \quad \textbf{(1)}$$

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	REGR factor score 1 for analysis 1	,544	,144	**,544**	3,770	**,001**	,245	,843	,544	,626	,544	**1,000**	**1,000**
	REGR factor score 4 for analysis 1	,343	,144	**,343**	2,380	**,026**	,044	,643	,343	,452	,343	**1,000**	**1,000**
	REGR factor score 5 for analysis 1	,358	,144	**,358**	2,482	**,021**	,059	,657	,358	,468	,358	**1,000**	**1,000**

a. Dependent Variable: Zscore: maximum expiratory pressure

b. Linear Regression through the Origin

Table 10: Standardized Coefficients of Regression Factors

After substituting C1, C4, C5, respectively in Eq. (1), we obtained the standardized linear regression equation, $\widehat{y}' = \sum b_i' X'_i$, Eq. (2)

$$\widehat{y}' = 0.50*X'_1 + 0.49*X'_2 + 0.50*X'_3 + 0.24*X'_4 + 0.59*X'_5 - 0.39*X'_6 - 0.48*X'_7 - 0.28*X'_8 \quad (2)$$

Subsequently, we are going to obtain the general linear regression equation with the original set of variables through the following procedure. In SPSS, select Analyze → Correlate → Bivariate, in Bivariate Correlations: Options we choose 'Cross-product deviations and covariances' and the Table 11 is generated.

Correlations

		age in years	height (cm)	weight (kg)	body mass (% of normal)	forced expiratory volume	residual volume	functional residual capacity	total lung capacity	maximum expiratory pressure
age in years	Pearson Correlation	1	,926**	,906**	,378	,294	-,532**	-,639**	-,469*	,613**
	Sig. (2-tailed)		,000	,000	,063	,153	,006	,001	,018	,001
	Sum of Squares and Cross-products	**614,240**	2417,400	1968,552	550,640	400,360	-5537,600	-3393,800	-967,000	2490,560
	Covariance	25,593	100,725	82,023	22,943	16,682	-230,733	-141,408	-40,292	103,773
	N	25	25	25	25	25	25	25	25	25
height (cm)	Pearson Correlation	,926**	1	,921**	,441*	,317	-,571**	-,624**	-,457*	,599**
	Sig. (2-tailed)	,000		,000	,027	,123	,003	,001	,022	,002
	Sum of Squares and Cross-products	2417,400	**11094,000**	8503,020	2730,400	1829,600	-25260,000	-14083,000	-4002,000	10338,600
	Covariance	100,725	462,250	354,293	113,767	76,233	-1052,500	-586,792	-166,750	430,775
	N	25	25	25	25	25	25	25	25	25
weight (kg)	Pearson Correlation	,906**	,921**	1	,673**	,449*	-,633**	-,617**	-,418*	,635**
	Sig. (2-tailed)	,000	,000		,000	,024	,001	,001	,037	,001
	Sum of Squares and Cross-products	1968,552	8503,020	**7688,230**	3468,272	2158,828	-23288,380	-11591,840	-3050,100	9123,688
	Covariance	82,023	354,293	320,343	144,511	89,951	-970,349	-482,993	-127,087	380,154
	N	25	25	25	25	25	25	25	25	25
body mass (% of normal)	Pearson Correlation	,378	,441*	,673**	1	,546**	-,615**	-,434*	-,365	,230
	Sig. (2-tailed)	,063	,027	,000		,005	,001	,030	,073	,270
	Sum of Squares and Cross-products	550,640	2730,400	3468,272	**3459,040**	1759,960	-15178,600	-5471,800	-1784,000	2211,160
	Covariance	22,943	113,767	144,511	144,127	73,332	-632,442	-227,992	-74,333	92,132
	N	25	25	25	25	25	25	25	25	25
forced expiratory volume	Pearson Correlation	,294	,317	,449*	,546**	1	-,691**	-,665**	-,443*	,453*
	Sig. (2-tailed)	,153	,123	,024	,005		,000	,000	,027	,023
	Sum of Squares and Cross-products	400,360	1829,600	2158,828	1759,960	**3009,040**	-15906,400	-7814,200	-2020,000	4073,840
	Covariance	16,682	76,233	89,951	73,332	125,377	-662,767	-325,592	-84,167	169,743
	N	25	25	25	25	25	25	25	25	25
residual volume	Pearson Correlation	-,532**	-,571**	-,633**	-,615**	-,691**	1	,877**	,607**	-,284
	Sig. (2-tailed)	,006	,003	,001	,001	,000		,000	,001	,168
	Sum of Squares and Cross-products	-5537,600	-25260,000	-23288,380	-15178,600	-15906,400	**176134,000**	78819,000	21177,000	-19542,400
	Covariance	-230,733	-1052,500	-970,349	-632,442	-662,767	7338,917	3284,125	882,375	-814,267
	N	25	25	25	25	25	25	25	25	25
functional residual capacity	Pearson Correlation	-,639**	-,624**	-,617**	-,434*	-,665**	,877**	1	,704**	-,417*
	Sig. (2-tailed)	,001	,001	,001	,030	,000	,000		,000	,038
	Sum of Squares and Cross-products	-3393,800	-14083,000	-11591,840	-5471,800	-7814,200	78819,000	**45872,000**	12541,000	-14637,200
	Covariance	-141,408	-586,792	-482,993	-227,992	-325,592	3284,125	1911,333	522,542	-609,883
	N	25	25	25	25	25	25	25	25	25
total lung capacity	Pearson Correlation	-,469*	-,457*	-,418*	-,365	-,443*	,607**	,704**	1	-,182
	Sig. (2-tailed)	,018	,022	,037	,073	,027	,001	,000		,385
	Sum of Squares and Cross-products	-967,000	-4002,000	-3050,100	-1784,000	-2020,000	21177,000	12541,000	**6910,000**	-2473,000
	Covariance	-40,292	-166,750	-127,087	-74,333	-84,167	882,375	522,542	287,917	-103,042
	N	25	25	25	25	25	25	25	25	25
maximum expiratory pressure	Pearson Correlation	,613**	,599**	,635**	,230	,453*	-,284	-,417*	-,182	1
	Sig. (2-tailed)	,001	,002	,001	,270	,023	,168	,038	,385	
	Sum of Squares and Cross-products	2490,560	10338,600	9123,688	2211,160	4073,840	-19542,400	-14637,200	-2473,000	**26832,640**
	Covariance	103,773	430,775	380,154	92,132	169,743	-814,267	-609,883	-103,042	1118,027
	N	25	25	25	25	25	25	25	25	25

Table 11: Sum of Squares and Cross-products

Partial regression coefficients and constant were computed as shown in Eq. (3), Eq. (4).

$$b_o = \bar{Y} - \sum b_i \bar{X}_i \quad (3)$$

where $\bar{Y}, \bar{X}_i$ were obtained from Table 5.

$$b_i = b_i' * \sqrt{\frac{Lyy}{Lx_i x_i}} \quad (4)$$

b_i: the *i*th partial regression coefficient of general linear regression equation

b_i': the *i*th partial standardized coefficient of standardized linear regression equation

Lyy: the Sum of Squares of the dependent variable Y

$Lx_i x_i$: Sum of Squares of the *i*th independent variable X_i

b_0: the constant of the general linear regression equation

where $Lyy, Lx_i x_i$ were obtained from Table 11 and are the following:

Lyy = 26832.640, Lx_1x_1 = 614.240, Lx_2x_2 = 11094.000, Lx_3x_3 = 7688.230, Lx_4x_4 = 3459.040, Lx_5x_5 = 3009.040, Lx_6x_6 = 176134.000, Lx_7x_7 = 45872.000, Lx_8x_8 = 6910.000

Finally, transforming the standardized linear equation, Eq. (2), into the general linear regression equation as shown in Eq. (5).

$$\hat{y} = b_o + \sum b_i X_i \quad (5)$$

$$\boldsymbol{\hat{y} = -43.9 + 3.30*X_1 + 0.76*X_2 + 0.94*X_3 + 0.66*X_4 + 1.76*X_5 - 0.15*X_6 - 0.36*X_7 - 0.56*X_8} \quad \boldsymbol{(5)}$$

Chapter 3: Conclusions

In this thesis, we manage to detect and address the problem of multicollinearity by applying Principal Component Analysis in this medical related data set, where Age, Height and Weight were the independent variables that are highly correlated. We end up with a new set of variables, the principal components, which are uncorrelated and which are ordered so that the first few retain the most of variation present in all of the original variables.[10] Principal Component Analysis is performed easily and effectively with SPSS statistical package.

In the literature, the adverse impact of ignoring multicollinearity is very well documented as Vatcheva et al. point out in the paper 'Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies'. They encourage researchers to proceed in Multicollinearity Diagnostics in regression analysis by reviewing epidemiological literature in PubMed from January 2004 to December 2013 and presenting the significance of detecting highly correlated predictors. The search that their team conducted reveals that in PubMed the terms collinearity, multicollinearity, collinear or multicollinear were found in only 0.12% of the studies that used multivariable regression. A percentage that it is presented skeptically since it was not clear if these studies had actually multicollinear data sets.[11]

In view of all that has been mentioned so far, it is important to consider multicollinearity diagnostics when analyzing data using regression models, avoiding misleading and erroneous interpretations of the results.

References

1. Chatterjee-Ali-S-Hadi, S. *Regression Analysis by example.*
2. Tu YK, Kellett M, Clerehugh V, Gilthorpe MS., Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. *British Dental Journal 199, 457 - 461 (2005) DOI:10.1038/sj.bdj.4812743*
3. Multicollinearity: Definition, Causes, Examples Retrieved from http://www.statisticshowto.com/
4. Batsidis Apostolos, 2017, *Biostatistics notes*
5. Xiao Sinan, Lu Zhenzhou, Xu Liyang, 2017, Multivariate sensitivity analysis based on the direction of eigen space through principal component analysis. *Reliability Engineering and System Safety 165 (2017) 1-10.*
6. Wikipedia, Retrieved from https://en.wikipedia.org/wiki/Principal_component_analysis
7. Pearson K. (1901). On lines and planes of closest fit to systems of points in space.
8. Julien Hoffman, *Biostatistics for Medical and Biomedical Practitioners* [https://www.elsevier.com/books/biostatistics-for-medical-and-biomedical-practitioners/hoffman/978-0-12-802387-7]
9. Liu R.X., Kuang J., Gong Q. Hou X.L., Principal component regression analysis with SPSS. *Computer Methods and Programs in Biomedicine, 2003; 71:141 - 147*
10. Jolliffe I.T., 2002, *Principal Component Analysis*
11. Kristina P. Vatcheva, MinJae Lee, Joseph B. McCormick, and Mohammad H. Rahbar, Epidemiology (Sunnyvale). 2016 April; 6(2), Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies doi:10.4172/2161-1165.1000227.